Displacing big data: How criminals cheat the system

Alice Hutchings[1], Sergio Pastrana[2], and Richard Clayton[1]

[1] Cambridge Cybercrime Centre, Department of Computer Science and Technology, University of Cambridge
[2] Department of Computer Science and Technology, University Carlos III of Madrid

Abstract. Many technical approaches for detecting and preventing cybercrime utilise big data and machine learning, drawing upon knowledge about the behaviour of legitimate customers and indicators of cybercrime. These include fraud detection systems, behavioural analysis, spam detection, intrusion detection systems, anti-virus software, and denial of service attack protection. However, criminals have adapted their methods in response to big data systems. We present case studies for a number of different cybercrime types to highlight the methods used for cheating such systems. We argue that big data solutions are not a silver bullet approach to disrupting cybercrime, but rather represent a Red Queen's race, requiring constant running to stay in one spot.

Keywords: Cybercrime; crime prevention; displacement; big data; machine learning

1 Introduction

In this chapter, we consider how commercial organisations are using 'big data' processing to detect and prevent cybercrime. Typically, a large dataset is created to hold information about how customers interact with the organisation. Known instances of crime are then considered, with the aim of determining what differences in behaviour can be identified when comparing criminals with all the legitimate customers. Going forward, the organisation will use these differences to flag up behaviour which matches the criminal profile with a view to preventing further crimes. While this type of big data approach is used by national security and policing agencies (Chan & Moses, 2017), we focus on established uses by financial institutions and technology companies.

Displacement occurs when offenders or crime changes as a direct result of preventative actions, and can result in crime changing and evolving (Cornish & Clarke, 1987). Online, offenders do not have to physically relocate in order to displace their activities. Smith et al. (2003) characterised the different types of displacement that can occur online. These including moving to new locations, times, targets, methods, offenders or offence types.

Displacement has been noted in relation to a number of cybercrime types as a direct result of prevention measures. For example, offenders providing denial of

service attacks for a fee (known as 'booter services') have come across a number of obstacles designed to stymie their activities, such as patching the servers being used to reflect traffic and blocking the accounts they use to receive payment. In these situations, offenders have displaced by changing the methods used for denial of service attacks (Hutchings & Clayton, 2015) and accepting new forms of payment such as bitcoin (Karami, Park, & McCoy, 2016).

Displacement is very evident following website takedown, a popular disruption method for a number of cybercrime types, most notably phishing (Moore & Clayton, 2007). Hutchings, Clayton and Anderson (2016) explored the displacement effects observed following website takedown. These include displacing to new domain names or new hosting providers, including 'bulletproof' and abuse tolerant providers. Offenders also selected different targets, such as different brands, and diversified phishing targets away from financial institutions towards other online services. New methods included technical approaches to make takedown harder. As well as displacing, websites are also simply replaced, reappearing in the same location. This typically occurs with malicious pages hosted within compromised websites, when the vulnerability that lead to the initial compromise is not fixed. As a result, takedown is often described as a 'whack-a-mole' approach, as websites tend to pop back up soon after (Hutchings et al., 2016; Chia, Chuang, & Chen, 2016).

Some crimes that are commonly detected using big data solutions, such as fraud detection systems, have also been seen to displace in response to intervention. For example, credit card fraud used to be the main method for fraudulently obtaining airline tickets. However, as detection of these unauthorised transactions improved, offenders changed their operations, including: compromising loyalty point accounts; phishing travel agencies for their access credentials for booking systems; and compromising business accounts (Hutchings, 2018b).

While these are relatively straightforward examples of displacement, in this chapter we instead focus on more complicated ways that criminals have designed ways of *circumventing* big data systems to avoid detection. We demonstrate the practical limitations of machine learning and big data approaches in adversarial settings. To do this, we explore how online crime and crime prevention techniques have co-evolved over time.

2 Cybercrime risks and big data approaches

The systems we describe here use algorithms to detect unusual activity that may point to active instances of particular types of cybercrime, namely use of compromised credit cards, access to compromised accounts, malicious communications, unauthorised access to computer systems, malware infection, and denial of service attacks.

Machine learning approaches may use one of a variety of techniques, including supervised, unsupervised, and reinforcement approaches (Robert, 2014). Supervised approaches require training data that have been pre-labelled. Unsupervised learning involves a fully automated approach without any pre-labelled training

data, with the aim of identifying interesting patterns. Reinforcement learning involves reward or punishment signals.

While many of the specifics of how these systems work are the 'secret sauce' of the commercial organisations that provide them, we do know that many use big data approaches, typically learning from what 'normal' behaviour looks like (and therefore detecting unusual behaviour), as well as 'red flags', or activities that are indicative of crime.

We start by describing the crimes and detection systems. We will later demonstrate how these mitigations can be circumvented by motivated offenders, leading to displacement and offence evolution.

2.1 Use of compromised credit cards

The risk: Credit cards can be compromised in a number of ways, both online and offline. Methods include data breaches, skimmers installed on point of sale terminals and ATMs, malware, phishing, theft, and physical mail interception. Depending on the methods by which it is obtained, credit card data can come in a variety of forms. 'Dumps' include the data read from the magnetic stripes on the back of cards, obtained by skimming, and can be used to create card clones. Credit card numbers, expiry dates, and card verification values, found on the back of the card, are required to process card-not-present payments. 'Fullz' refers to the full information associated with the account, including data relating to the account holder, such as name, address, and date of birth (Hutchings & Holt, 2015).

Stolen data markets provide platforms for the sale of compromised credit card data (Franklin, Paxson, Perrig, & Savage, 2007; Holt & Lampke, 2010; Hutchings & Holt, 2015; Motoyama, McCoy, Levchenko, Savage, & Voelker, 2011). Cloned and stolen cards can be used over the counter, or at ATMs if PINs are also obtained. However, card-not-present transactions require data that can be obtained without physical contact with the card, and can be completed online, by phone or by mail order. The main objective is to monetise the compromised cards, such as by selling on purchased goods and services (Hutchings & Holt, 2015).

Fraud detection systems: Fraud detection systems are used to detect credit card fraud at the time of the transaction (Abdallah, Maarof, & Zainal, 2016; Phua, Lee, Smith, & Gayler, 2010). These systems use algorithms to score the potential risk. They identify patterns that do not match the cardholder, such as the IP address and location, address provided, browser, language setting, and spending patterns. They also identify patterns that match known frauds, such as the type of purchase and associated variables. A risky transaction can then be reviewed, and attempts can be made to confirm its legitimacy by making enquiries with the cardholder. If there is confirmed fraud, or significant doubt, the transaction may be cancelled, or another payment method requested.

Fraud detection systems may be used by the merchant, as well as by financial institutions attempting to detect fraudulent transactions on their cards. Third

party vendors and payment processors may also have their own fraud detection systems. Multiple parties operating their own fraud detection systems is beneficial, as while one party, such as a merchant, may see deeply into fraud on their own systems, another party, such as a bank or third party vendor, may see more breadth, with attempted transactions across multiple targets (Hutchings, 2018b).

2.2 Access to compromised accounts

The risk: There are a variety of accounts that may be compromised, such as bank, email, social media, merchant, and gaming accounts. There are also a number of ways accounts can be compromised, either at scale or by targeting particular account holders. For example, data breaches may provide access to large numbers of online accounts. Furthermore, if username and password combinations are breached, these may be checked against other online systems to gain access to accounts where credentials have been reused. Therefore, accounts may be compromised due to opportunity, as well as through targeted attacks. Accounts may also be compromised through the use of malware or phishing (Onaolapo, Mariconti, & Stringhini, 2016). Account recovery systems can also be misused to gain access to particular accounts, particularly where 'secret' questions can be inferred through other public or known information.

Some account credentials are offered for sale in online criminal markets, while others may be released publicly on websites such as Pastebin (Onaolapo et al., 2016). However, accounts can also be used by those that obtained them, or their accomplices. Accounts can be used to commit other types of offences, such as email and social media accounts for the purpose of sending malicious communications (as detailed in the following section).

Onaolapo et al. (2016) monitored the way compromised Gmail accounts are used by purposely leaking accounts under their control and observing the subsequent activity. Most visitors appeared to be simply curious, and did not perform any further actions after accessing the accounts. However, other visitors using the leaked credentials searched for potentially sensitive information that could be monetised, or they sent email spam, and some attempted to lock out the account holder by changing the password.

Behavioural analysis: The Messaging, Malware and Mobile Anti-Abuse Working Group (M^3AAWG) is an industry group that develops cooperative approaches for the purpose of combating cybercrime. Among their 'best practice' guides, M^3AAWG (2014) have recommendations for detecting the types of unusual activity associated with compromised accounts. They recommend using systems that require big data solutions, which can then detect unusual account activity, such as access from different locations or devices. While this analytic approach compares activity to the usual behaviour of the account holder, other approaches compare behaviours with the activity that typically occurs after compromise.

Indications that an account may be compromised may differ by account type. For example, for email accounts, indications could include sending messages

to all contacts, or the deletion of sent mail. For some other account types, a typical pattern associated with compromise is immediate steps to lock out the account holder. Changing passwords and associated email addresses rarely occurs spontaneously, which facilitates detection.

2.3 Malicious communications

The risk: As identified above, some accounts may be compromised for the purpose of sending malicious communications. Other accounts may be specifically set up for this purpose. Malicious communications can be sent on many different platforms, and can take the form of email, chat, text messages, or social media posts. The malicious purpose of the communication may include disseminating spam or phishing URLs, distributing malware, or attempts at social engineering for fraudulent purposes.

Moore and Clayton (2015) analyse how malicious communications have spread across instant messaging and social media systems. The messages analysed contained URLs, and when the website was visited by the recipient, they could also be infected with malware, and those in their contact lists were in turn sent a copy of the malicious message. By monitoring the command and control (C&C) channel that was issuing instructions to the malware, they found that offenders changed their methods in response to efforts made to take down websites that the URLs directed to, such as using URL-shorteners for malicious links.

Spam detection: One way to detect malicious communications is to measure behaviour that is different to the purported sender (Egele, Stringhini, Kruegel, & Vigna, 2017). In addition to the behavioural analyses that may indicate that an account has been compromised (as discussed above), anomalous patterns can be identified in the messages that are sent, and the associated metadata.

Patterns can also be detected for similar messages to known malicious communications. For example, for malicious messages sent over social media, this could include the message content, and the presence of suspicious URLs (Egele et al., 2017). Additional factors may include attachments, and for email, 'spoofing' (a term used in computer security to refer to events in which an attacker masquerades as another party) the header information, to make it appear it had been sent by a legitimate organisation (Fette, Sadeh, & Tomasic, 2007).

2.4 Unauthorised access to computer systems

The risk: Remote exploitation mechanisms rely on system vulnerabilities to gain unauthorised access to a computer system. Exploitation usually consists of three phases: scanning the network, sending the exploit, and post-exploitation. Prior to triggering an exploit, offenders usually scan networks (or the whole Internet) to find vulnerable machines. Once a vulnerability is discovered, a crafted network packet containing the corresponding exploit is sent. The post-exploitation phase usually involves the compromised system connecting to a

remote server (also known as command and control, or C&C server) for further instructions, for example, to download additional malware which will persist after the machine reboots.

Another way of gaining unauthorised access to a machine is by means of escalation of privileges, where users with restricted permissions gain access to unauthorised assets within the same system or network. Vulnerable machines, weak passwords and weak security policies (e.g. those that do not enforce access control or have configuration flaws) are the most common causes that allow for escalation of privileges (Affinity IT Security Services, 2017).

Intrusion detection systems: Intrusion detection systems (IDS) are one of the oldest security mechanisms used to protect systems and networks (Denning, 1987). These systems look for patterns of malicious behaviour in either the network (Network IDS, or NIDS) or the host activity (Host IDS or HIDS). An IDS will trigger alarms when any suspicious activity is monitored. Prior to exposing the alerts to human operators for further inspection, IDS alarms might be further correlated with other activity gathered from the systems or networks, such as system or router logs, or even other IDS alarms in Security Information and Event Management (SIEM) systems.

IDSs are typically classified based on their detection mode. *Anomaly-based* IDS compute a model of normal behaviour and trigger alarms when they monitor activity that does not fit within the model. *Signature-based* IDS analyse the monitored activity looking for malicious patterns which are encoded in a predefined set of rules or 'signatures'. The main challenge for an anomaly-based IDS is to compute a model that represents faithfully the normality, due to the current complexity of current systems and networks. This paradigm leads to a higher false positive rate, i.e. normal activity being tagged as suspicious. On the contrary, signature-based IDS are more precise when detecting known attacks, but are ineffective at detecting so called zero-day exploits, i.e. attacks that have not been seen previously and for which there are no known signatures.

The use of big data technologies for intrusion detection systems is widespread. This is mainly due to the large number of network packets and system events that need to be processed. Example analytical approaches include machine learning, graph analysis, and clustering, which have previously been applied to characterise traffic sent to and from C&C servers (Gardiner & Nagaraja, 2016). Additionally, the increased use of 'Internet of Things' devices and mobile sensors pose additional challenges which require big data approaches. Previous studies showed that intrusion detection for Mobile Ad-Hoc Networks (MANETs) can be performed by means of evolutionary computation techniques (Sen & Clark, 2011) or machine learning classifiers (Pastrana, Mitrokotsa, Orfila, & Peris-Lopez, 2012).

2.5 Malware infection

The risk: Malware, or malicious software, poses a number of risks. Malware infections can result in stolen data and compromised credentials (Hutchings &

Clayton, 2017). Malware can create botnets, whereby the connected machines can be controlled to perform acts in concert, such as denial of service attacks or phishing campaigns. Malware known as 'ransomware' encrypts data, demanding payment of a ransom if the victim wants to regain access. Attackers using malware to compromise victims' machines may have additional motivations, such as accessing the webcam for voyeurism.

Anti-virus software: Anti-virus programs constantly run in the background of systems looking for evidence of malware infection. Commonly, anti-virus companies manage a list of known bad applications, such as fingerprints of malicious files, blacklists of IP addresses, or malicious programs that are known to be loaded by malware. These lists must be periodically updated by the anti-virus program from the anti-virus company server to prevent infection from new pieces of malware.

Anti-virus companies receive a large number of new samples to analyse every day. There are two approaches for binary analysis: static and dynamic analysis. Static analysis focuses on the binary itself, analysing the contents without executing it to detect suspicious activity (i.e. without installing and running the software). Dynamic analysis is performed by executing the malware in a 'sandbox'. The term sandbox refers to a controlled machine where the binary can be safely executed and its behaviour can be analysed (Gandotra, Bansal, & Sofat, 2014).

According to Panda Labs (2017), up to 285,000 suspected malware programs were analysed daily during 2017.[3] Given this threat scenario, big data approaches are necessary to rapidly classify the analysed binaries. There is a large volume of research applying machine learning for the purpose of classifying malware samples. For example, Rieck et al. (2011) applied clustering algorithms to group malware samples into families and detect botnet campaigns based on the responses from the Domain Name System (DNS) server. Dash et al. (2016) used machine learning to classify Android malware into families, based on the activity monitored during the execution of the samples in a controlled environment.

2.6 Denial of service attacks

The risk: Distributed denial of service (DDoS) attacks involve overloading a website or computer system with bogus traffic, thereby blocking legitimate access. DDoS attacks have a wide range of targets, such as corporations, governments, and gaming servers (Karami & McCoy, 2013). The targets may vary with the purpose of the attack, such as extortion (demanding payment for 'protection' against further attacks), and protest. Booter services provide DDoS attacks as a service, which is primarily advertised towards gamers, offering them an advantage against their adversaries (Hutchings & Clayton, 2015).

[3] We note that statistical data from commercial organisations should be analysed critically (Anderson et al., 2013).

At one time, DDoS attacks were primarily carried out using botnets, following malware infection. However, 'amplification' or 'reflection' methods allow attackers to have greater power with limited resources. Amplification attacks involve spoofing the victim's IP address, and sending a query to another server. As it appears the question was sent by the victim, the response is returned to them. Since the responses are larger than the requests the attacker has amplified the amount of traffic sent to the victim, reflected from another server, compared to the resources they actually deployed (Thomas, Clayton, & Beresford, 2017).

Denial of service attack protection: There is now a substantial industry providing DDoS mitigation services. The basic idea is to place a device 'in front of' the systems to be protected and inspect the incoming traffic. Malicious traffic is discarded but 'good' traffic is passed on the service which can then respond as normal. This will keep the service available for legitimate users provided the detection is accurate and provided that the malicious traffic does not exhaust the available bandwidth. Complex arrangements to reroute traffic are used to deal with bandwidth issues, and indeed just to arrange that the filtering device can be put 'in front of' the system to be protected. Further information about the type of systems involved can by found in the survey by Zargar et al. (2013) and the book by Yu (2014).

Filtering systems originally depended on simple heuristics or used custom filters specially created by a human to deal with the particular attack. Recently there has been considerable academic work done on machine learning systems that use a wide range of traffic characteristics to determine whether or not traffic is malicious (and whether or not there is an attack going on at all) (Mayhew, Atighetchi, Adler, & Greenstadt, 2015; Sommer & Paxson, 2010; Zuech, Khoshgoftaar, & Wald, 2015). However, the actual devices are made by commercial companies and they provide no details of their technology.

3 Cheating the system

During the course of our research careers, we have identified a number of ways criminals have cheated big data systems, in order to circumvent detection. In some of the case studies we outline below, we acknowledge we are deliberately vague about certain details, in order to protect specific organisations and industry types, and to avoid further malevolent development of these approaches.

3.1 Loopholes

Fraud detection systems are not foolproof. In particular, they have mainly been developed to detect fraudulent credit card transactions, so few will detect more elaborate frauds. Many merchants use fraud detection systems within their online shops where there is a wealth of data that can inform fraud risk, such as IP addresses and device fingerprints. However, some merchants may also operate

call centres that do not use such systems, and even if they did, there is far less data to go on.

Some types of transactions are time-sensitive. However, confirming fraud can take time. When fraud is identified or suspected, merchants will usually check with the bank to verify a transaction is not authorised. This introduces delays, especially when crossing international borders, as there can be timezone and language differences. This issue is compounded when there is no incentive for the issuing bank to detect fraud themselves, or respond in a timely fashion. If the purchase is fraudulent they suffer no financial loss, with the merchant generally liable for the chargeback for card-not-present transactions.

3.2 Imitation

Tutorials are available on underground forums and markets which teach others how to circumvent fraud detection systems. Tutorials found on stolen data markets by Hutchings and Holt (2015) detailed how to imitate genuine cardholders, including: using virtual machines to change the operating system; changing settings, such as language and timezone, for operating systems and browsers; and changing the IP address to one near the victim's usual residence, by using proxies, anonymity networks such as Tor, or virtual private networks. An example extract from such a tutorial, translated from Russian, reads:

> *very simple while working you have to fU....k the security system and for this reason your comp must be American and namely the language, time and even the username*
> *everything should be like it is with a real americosa.*
> *in order to reduce suspicion against you to a more acceptable level take this seriously here every detail is important, the winda language, socks, browser language even the time must be completely set for the state in wich we are working. (In certain situations it is necessary to pay attention to what's time it is nooww in the cardholder's country and not type too late or too early)*

Other research has found support for the idea that offenders will imitate genuine account holders. Onaolapo et al. (2016) found that when the location of leaked account holders was known, some offenders would connect to the account using IP addresses in nearby locations. Others disguised their browser 'user agent' string, which provides a website with information about the visitor's browser and operating system.

Another way to imitate a user is through the use of cookies. Cookies are small pieces of data stored in the web browser. Some malware will steal users' cookies. These can be used to fool websites into treating an attacker as a logged in user (Hutchings & Clayton, 2017).

Future ways to imitate include building accurate video and audio models of targets. Some organisations are investing in voice and facial recognition systems in order to positively identify customers. However, in the future attackers

may use the same types of data that are used to train recognition systems for impersonation of selected targets (Riek & Watson, 2010).

3.3 Appear innocuous

Ways to appear innocuous include changing what is being purchased from something that is not suspicious, but the offender does not really want, to the real target after the fraud detection system checks have been completed. This technique can be used for transactions that are likely to be flagged as high risk. The offender first completes a transaction that does not raise a red flag. Once the order has been confirmed, they can then contact the merchant, still posing as the genuine cardholder, and change the order to reflect what they actually do want.

Another way of avoiding red flags is to tailor the order to make it appear less suspicious. Fraudulently obtained airline tickets tend to be booked shortly before departure, so as to limit the likelihood that the transaction will be flagged before the flight departs. This resulted in one-way flights being booked for each route, with bookings for return or subsequent destinations being made separately. However, one-way flights then became a red flag for flights. Subsequently, Hutchings (2018b) found there is often a return flight booked, even if there is no intention of flying it. As the booking is made using fraudulent means, the second flight doesn't come at a cost to the traveller.

In relation to malware, there are techniques to make the software appear innocuous in the eyes of anti-virus signatures. Packing refers to compressing malware executables in order to obscure their contents. This makes it harder for anti-virus software to detect the malware. However, packers can be detected and some anti-virus software will flag everything that has been packed. More advanced techniques are possible, such as those that change the morphology of the binaries so they are not detected, without encrypting the entire file (Kruegel, Kirda, Mutz, Robertson, & Vigna, 2005).

A common approach to evade intrusion detection systems requires acquiring knowledge about how it works. This can be done by means of probe attacks, where the attacker queries the IDSs and analyse their responses (Pastrana, Orfila, & Ribagorda, 2011). This information allow offenders to create exploits that mimic regular network traffic and bypass detection (Vigna, Robertson, & Balzarotti, 2004; Fogla, Sharif, Perdisci, Kolesnikov, & Lee, 2006; Pastrana, Orfila, & Ribagorda, 2010). Some researchers have proposed the use of random detection functions to combat these mimicry attacks (Wang, Parekh, & Stolfo, 2006). However, such randomised schemes are still vulnerable if the adversary is able to interact with the detector for a longer period of time and infer randomization patterns (Pastrana, Orfila, Tapiador, & Peris-Lopez, 2014).

3.4 Insiders

The use of insiders can help offenders circumvent fraud detection systems entirely, or to learn how to reverse engineer and avoid them. Offenders may seek

to obtain employment at targeted organisations. On one stolen data market, Hutchings and Holt (2015) found suggestions that specialist knowledge could be learnt by applying for employment at a company offering fraud detection systems. There were indications that the news and developments relating to organisations of interest were being closely followed for the purpose of seeking such opportunities. For example, the following post related to a consumer credit reporting agency, which was moving offices and advertising for new staff:

> *Another change that will be occuring within* [company] *is the closeure of all of their regional offices.* [The company] *is going to consolidate to one fixed location in* [city]. *The reason they are doing this is, and I quote, "To lover overhead costs, and also to increase security." So don't be too overly suprised if* [the company] *seems to be learning a few new tricks. The plan is to be completed within 2–3 years, with most of the outlying sites already closed. One nice note to this... If you live in* [city], *look in the want adds,* [the company is] *hiring Data Entry personnel... Hmmmm...*

Offenders may also target employees and contractors, to either corrupt or blackmail them into providing assistance. In another post, it was suggested that employees with knowledge of fraud detection systems and 'morals that are questionable' should be targeted, in order to learn about how to avoid detection:

> *Right now I would like to get a hold of the* [...] *program that all the check cashing places use. With this we could figure out EXACTLY how there system works, and there is NEVER a callback with the program. If anyone works for a place that uses this software, or knows someone who has morals which are questionable, please contact me. I will make it very worthwhile for you to do so...*

3.5 Target the unprotected

Larger organisations will have significant resources available to implement systems that will reduce the amount of crime they experience. They benefit from economies of scale, due to the amount of trade that they do. However, smaller companies are less likely to be able to pay for fraud detection systems, and there are indications that fraudsters know this. When studying the trade in fraudulently obtained airline tickets, Hutchings (2018b) found that offenders recommended purchasing tickets from smaller travel agencies to avoid detection:

> *its better to go through smaller companies that cant pay for the extra fraud detection*

Unfortunately, smaller companies are also less likely to be able to withstand the losses arising from fraud, particularly if they are repeatedly targeted within a short period of time. Some small family-run companies have gone out of business for this reason (Hutchings, 2018a), contributing to the monopolisation of trade, and shutting down competition (particularly independent organisations).

3.6 Adversarial machine learning

Many of the crime detection techniques rely on a model which is constructed using a training dataset which informs the machine learning algorithms what should be considered 'good' or 'bad'. In the case of anomaly detection, this model represents what is normal, and the algorithm aims to detect outliers. In the case of signature-based detection, this model is a classifier of events into either malicious or regular events (spam vs regular email, intrusion vs normal traffic, malware vs benign software etc.). In recent years, researchers have identified a weakness in the use of machine learning under adversarial scenarios. Since these algorithms have not been designed with security in mind, a sophisticated adversary might be able to cheat the system.

A seminal paper by Barreno et al. questioned for the first time the security of machine learning (Barreno, Nelson, Sears, Joseph, & Tygar, 2006), leading to a number of researchers working on this problem. Huang et al. (2011) presented a taxonomy of potential attacks against machine learning algorithms with three main aspects: the influence (causative, if the adversary targets ('poisons') the training data, or exploratory, if the attacker targets the system once it is trained), the security violation (availability, integrity or privacy of the data) and the specificity (targeted or indiscriminate). Biggio, Fumera, and Roli (2014) enriched this taxonomy by adding the set of adversarial capabilities regarding its knowledge about the algorithm or capabilities to modify the system both before and after it is trained.

In the literature there are examples on how adversaries can bypass spam filters (Biggio, Fumera, & Roli, 2014) or malware classifiers (Biggio, Rieck, et al., 2014) that make use of machine learning. In both cases, the adversary is able to evade the classifier by first acquiring knowledge about how it works and then a very small number of modifications to the attacks so as to bypass detection (e.g, by adding or removing specific words from spam messages). However, in the real world it may be rather more complicated. Miscreants may find it difficult to obtain sufficient knowledge about the system and its capabilities and may be unable to run enough tests to scope out its training and detection processes, so they will not be able to trick the detectors.

4 Discussion and conclusion

Big data solutions that aim to detect cybercrime rely on humans behaving in relatively predictive way. Attempts to commit cybercrime will tend to generate distinctive patterns, and so the good and the bad can each be identified. Offenders will try and cheat the system by trying to make their behaviour blend in, but the edge for big data solutions is that the business employing them is in a position to know far more about the behaviour of its customers than the bad guys ever will.

One of the key features of these big data solutions is the use of continuous feedback so that the systems learn and adapt in response to known criminal

activity. However, the systems continually need to know the 'ground truth', and it is necessary for a certain amount of malicious activity to be detected by the system or by other monitoring, so the system can track and adapt accordingly.

This feedback is why reporting spam leads to improved spam detection, the reports are used to train the system as to what is or is not spam. However, this means that the definition of spam is no longer 'bulk unsolicited email' but becomes 'email the user does not want in their inbox today'. This leads to two problems. First, the bad guys know the feedback votes matter, so they submit incorrect votes about their own emails. The response is to build another machine learning system to identify false votes and eliminate them before training the main system. Secondly, it is essential to whitelist particular types of mail, such as boarding passes, electricity bills, and so forth, otherwise a small number of people reporting these as spam will cause the machine learning system to treat these as spam for everybody.

These big data solutions are treading a delicate balance between false positives and true negatives. For example, while spam detectors want to ensure account holders do not receive unwanted emails, they also want to ensure that the email account holders do want to receive gets through. Fraud detection systems want to ensure that fraudulent transactions are blocked, but also that genuine transactions are quickly processed. If genuine transactions are blocked, or fraudulent ones are processed, there is not only immediate financial loss, but increased expenditure in staff time and more annoyed customers. If switching costs are low, customers may go elsewhere in the future.

Offenders are evidently learning techniques to circumvent big data solutions. As we show in this chapter, some of this learning is facilitated by discussions that occur on forums and marketplaces. There is also self-learning, through trial and error, reverse engineering the systems to identify what methods work and what does not. Resources are available online for individuals to learn cybersecurity skills for defensive purposes, however technically inclined offenders can also avail themselves to these opportunities. As identified, insiders provide another way to learn about the algorithms powering these systems.

However, the methods used for cheating the system can eventually be identified, either by the machine learning algorithms themselves, or by those that operate them identifying that criminal activity has been overlooked. The solution is to re-engineer the system and retrain it, whilst ensuring that the quality of the results is maintained. Software packing hides the malicious payload of an executable program, but once the system learns to identify packers it will rapidly learn to treat them as a red flag. But here again, there is a delicate balance. Offenders may learn that if they buy time-sensitive orders at the last minute, they can avoid the process required to confirm if a transaction is genuine. However, of buying at the last minute becomes a red flag then the organisation will have to weigh up the cost of inconvenience to genuine customers with the potential fraud risk. It may be more profitable to allow some fraud to occur rather than turn away genuine customers.

Financial incentives play an important part in everything we have discussed. For example, merchants usually carry the cost of fraud for card-not-present transactions, such as those that occur online. If they suspect a transaction is fraudulent, the normal process is to verify this with the cardholder's financial institution. To the financial institution, this is often not a priority (as they won't carry the cost of the transaction, the merchant will). Hence, verification can take time, particularly if the cardholder is not immediately available, and if there are communication barriers, such as different languages and timezones.

Finally, the most important thing to understand is that the use of machine learning against adversaries is quite unlike the use of machine learning for other types of prediction, such as the weather. Thunderclouds do not cheat and change their behaviour just because you have worked out where they are. In Lewis Carroll's book *Through the Looking Glass*, the Red Queen tells Alice "Now, here, you see, it takes all the running you can do, to keep in the same place. If you want to get somewhere else, you must run at least twice as fast as that!". These big data approaches represent a Red Queen's race. Running non-stop is required to stay in one spot.

Funding

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) [grant EP/M020320/1] for the University of Cambridge, Cambridge Cybercrime Centre. The opinions, findings, and conclusions or recommendations expressed are those of the authors and do not reflect those of the aforementioned funder.

Acknowledgements

We thank our colleagues at the Cambridge Cybercrime Centre, in particular Dr Daniel Thomas for his insightful feedback and comments. We also thank the anonymous reviewers for their helpful suggestions and advice.

References

Abdallah, A., Maarof, M. A., & Zainal, A. (2016). Fraud detection system: A survey. *Journal of Network and Computer Applications*, *68*, 90–113.

Affinity IT Security Services. (2017). *What is privilege escalation?* Retrieved 2018-06-14, from `https://perma.cc/MAP3-HUXK`

Anderson, R., Barton, C., Böhme, R., Clayton, R., Van Eeten, M. J., Levi, M., ... Savage, S. (2013). Measuring the cost of cybercrime. In *The Economics of Information Security and Privacy* (pp. 265–300). Springer.

Barreno, M., Nelson, B., Sears, R., Joseph, A. D., & Tygar, J. D. (2006). Can machine learning be secure? In *Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security* (pp. 16–25).

Biggio, B., Fumera, G., & Roli, F. (2014). Security evaluation of pattern classifiers under attack. *IEEE Transactions on Knowledge and Data Engineering*, *26*(4), 984–996.

Biggio, B., Rieck, K., Ariu, D., Wressnegger, C., Corona, I., Giacinto, G., & Roli, F. (2014). Poisoning behavioral malware clustering. In *Proceedings of the 2014 Workshop on Artificial Intelligent and Security* (pp. 27–36).

Chan, J., & Moses, L. B. (2017). Making sense of big data for security. *British Journal of Criminology*, *57*(2), 299-319.

Chia, P. H., Chuang, J., & Chen, Y. (2016). Whack-a-mole: Asymmetric conflict and guerrilla warfare in web security. In *Proceedings of the 15th Annual Workshop on the Economics of Information Security.*

Cornish, D. B., & Clarke, R. V. (1987). Understanding crime displacement: An application of rational choice theory. *Criminology*, *25*(4), 933-947.

Dash, S. K., Suarez-Tangil, G., Khan, S., Tam, K., Ahmadi, M., Kinder, J., & Cavallaro, L. (2016). Droidscribe: Classifying Android malware based on runtime behavior. In *Security and Privacy Workshops (SPW), IEEE* (pp. 252–261).

Denning, D. E. (1987). An intrusion-detection model. *IEEE Transactions on Software Engineering*, *2*, 222–232.

Egele, M., Stringhini, G., Kruegel, C., & Vigna, G. (2017). Towards detecting compromised accounts on social networks. *IEEE Transactions on Dependable and Secure Computing*, *14*(4), 447-460.

Fette, I., Sadeh, N., & Tomasic, A. (2007). Learning to detect phishing emails. In *Proceedings of the 16th International Conference on World Wide Web* (pp. 649–656). New York, NY, USA: ACM.

Fogla, P., Sharif, M. I., Perdisci, R., Kolesnikov, O. M., & Lee, W. (2006). Polymorphic blending attacks. In *USENIX Security Symposium* (pp. 241–256).

Franklin, J., Paxson, V., Perrig, A., & Savage, S. (2007). An inquiry into the nature and causes of the wealth of internet miscreants. *ACM Conference on Computer and Communications Security (CCS)*, Alexandria, October 29–November 2.

Gandotra, E., Bansal, D., & Sofat, S. (2014). Malware analysis and classification: A survey. *Journal of Information Security*, *5*(02), 56.

Gardiner, J., & Nagaraja, S. (2016). On the security of machine learning in malware C&C detection: A survey. *ACM Computing Surveys (CSUR)*, *49*(3), 59.

Holt, T. J., & Lampke, E. (2010). Exploring stolen data markets online: Products and market forces. *Criminal Justice Studies*, *23*(1), 33–50.

Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B. I., & Tygar, J. (2011). Adversarial machine learning. In *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence* (pp. 43–58).

Hutchings, A. (2018a). Flying in cyberspace: Policing global travel fraud. *Policing: A Journal of Policy and Practice*, advanced access. doi: 10.1093/police/pay063

Hutchings, A. (2018b). Leaving on a jet plane: The trade in fraudulently obtained airline tickets. *Crime, Law and Social Change*, *70*(4), 461–487.

Hutchings, A., & Clayton, R. (2015). Exploring the provision of online booter services. *Deviant Behaviour*, *37*(10), 1163-1178.

Hutchings, A., & Clayton, R. (2017). Configuring Zeus: A case study of online crime target selection and knowledge transmission. *APWG Symposium on Electronic Crime Research (eCrime)*, Arizona, April 25-27.

Hutchings, A., Clayton, R., & Anderson, R. (2016). Taking down websites to prevent crime. *APWG Symposium on Electronic Crime Research (eCrime)*, Toronto, June 1-3.

Hutchings, A., & Holt, T. J. (2015). A crime script analysis of the online stolen data market. *British Journal of Criminology*, *55*(3), 596-614.

Karami, M., & McCoy, D. (2013). Understanding the emerging threat of DDoS-as-a-Service. In *Presented as part of the 6th USENIX Workshop on Large-Scale Exploits and Emergent Threats.* Washington, D.C.: USENIX.

Karami, M., Park, Y., & McCoy, D. (2016). Stress testing the booters: Understanding and undermining the business of DDoS services. *International World Wide Web Conference (IW3C2)*, Quebec, April 11-15.

Kruegel, C., Kirda, E., Mutz, D., Robertson, W., & Vigna, G. (2005). Automating mimicry attacks using static binary analysis. In *Proceedings of the 14th Conference on USENIX Security Symposium* (Vol. 14, pp. 11–11). Baltimore, MD, USA: USENIX.

Mayhew, M. J., Atighetchi, M., Adler, A., & Greenstadt, R. (2015). Use of machine learning in big data analytics for insider threat detection. In *MILCOM* (pp. 915–922). IEEE.

Messaging, Malware and Mobile Anti-Abuse Working Group. (2014). *M^3AAWG Compromised User ID Best Practices.* Retrieved from `https://perma.cc/Z2PL-BQUS`

Moore, T., & Clayton, R. (2007). Examining the impact of website take-down on phishing. *APWG 2nd Annual eCrime Researchers Summit*, Pittsburgh, October 4-5.

Moore, T., & Clayton, R. (2015). Which malware lures work best? Measurements from a large instant messaging worm. *APWG Symposium on Electronic Crime Research (eCrime)*, Berlin.

Motoyama, M., McCoy, D., Levchenko, K., Savage, S., & Voelker, G. M. (2011). An analysis of underground forums. *ACM SIGCOMM Conference on Internet Measurement*, Barcelona, May 26-29.

Onaolapo, J., Mariconti, E., & Stringhini, G. (2016). What happens after you are pwnd: Understanding the use of leaked webmail credentials in the wild. In *Proceedings of the 2016 Internet Measurement Conference* (pp. 65–79). New York, NY, USA: ACM.

Panda Labs. (2017). *2017 in figures: The exponential growth of malware.* Retrieved 2018-05-23, from `https://perma.cc/R5FG-73YR`

Pastrana, S., Mitrokotsa, A., Orfila, A., & Peris-Lopez, P. (2012). Evaluation of classification algorithms for intrusion detection in MANETs. *Knowledge-*

Based Systems, *36*(0), 217 - 225.

Pastrana, S., Orfila, A., & Ribagorda, A. (2010). Modeling NIDS evasion with genetic programming. In *Worldcomp 2010: Security and Management* (pp. 444–448).

Pastrana, S., Orfila, A., & Ribagorda, A. (2011). A functional framework to evade network IDS. In *Hawaii International Conference on System Sciences (HICSS'11)* (p. 1-10). Koloa, Hawaii, USA: IEEE.

Pastrana, S., Orfila, A., Tapiador, J. E., & Peris-Lopez, P. (2014). Randomized anagram revisited. *Journal of Network and Computer Applications*, *41*, 182–196.

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*.

Rieck, K., Trinius, P., Willems, C., & Holz, T. (2011). Automatic analysis of malware behavior using machine learning. *Journal of Computer Security*, *19*(4), 639–668.

Riek, L. D., & Watson, R. N. M. (2010). The age of avatar realism: When seeing shouldn't be believing. *IEEE Robotics & Automation Magazine*, *17*(4), 37-42.

Robert, C. (2014). *Machine learning, a probabilistic perspective.* Taylor & Francis.

Sen, S., & Clark, J. A. (2011). Evolutionary computation techniques for intrusion detection in mobile ad hoc networks. *Computer Networks*, *55*(15), 3441–3457.

Smith, R. G., Wolanin, N., & Worthington, G. (2003). *Trends & Issues in Crime and Criminal Justice No. 243: e-Crime solutions and crime displacement.* Canberra: Australian Institute of Criminology.

Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. In *IEEE Symposium on Security and Privacy* (pp. 305–316). IEEE Computer Society.

Thomas, D. R., Clayton, R., & Beresford, A. R. (2017). 1000 days of UDP amplification DDoS attacks. In *APWG Symposium on Electronic Crime Research (eCrime)* (p. 79-84).

Vigna, G., Robertson, W., & Balzarotti, D. (2004, October). Testing network-based intrusion detection signatures using mutant exploits. In *Proceedings of the 11th ACM Conference on Computer and Communications Security* (p. 21). Washington, DC, USA: ACM.

Wang, K., Parekh, J. J., & Stolfo, S. J. (2006). Anagram: A content anomaly detector resistant to mimicry attack. In *International Workshop on Recent Advances in Intrusion Detection* (pp. 226–248).

Yu, S. (2014). *Distributed denial of service attack and defense.* Springer Publishing Company, Incorporated.

Zargar, S. T., Joshi, J., & Tipper, D. (2013). A survey of defense mechanisms against distributed denial of service (DDoS) flooding attacks. *IEEE Communications Surveys and Tutorials*, *15*(4), 2046-2069.

Zuech, R., Khoshgoftaar, T. M., & Wald, R. (2015, Feb 27). Intrusion detection and big heterogeneous data: a survey. *Journal of Big Data*, *2*(1), 1–41.